

**本科专业课程考试报告**

**Course Exam Report OF UNDERGRADUATE**

**(2025届)**

**大数据可视化实验报告**

房贷可视化实验

|  |  |
| --- | --- |
| **学　　院** | 光电信息与计算机工程学院 |
| **专　　业** | 数据科学与大数据技术 |
| **学生姓名** | 罗震宇 |
| **学　　号** | 2135060321 |
| **指导教师** | 艾均 |
| **成 绩** |  |

# 摘 要

本实验旨在通过可视化和可视分析技术，以及网络建模方法，从一本喜爱的图书中挖掘出有价值的信息。通过对文本的分词、虚词过滤、同义名词预处理等操作，设计了算法并用Python实现了复杂网络模型的提取。利用NetworkX包构建了网络模型，采用Gephi和NetworkX计算了网络的中心性和社团结构，并且使用Excel进行了可视化。以《朝闻道》小说为例，通过文本处理、网络建模和可视化，展示了实验的过程和结果。通过对小说网络拓扑结构的分析，验证了实际网络中社团内部联系较紧密、社团之间联系相对稀疏的论点。

通过本次实验，我对于将文本进行可视化的技能有了进一步的了解，同时Python编程能力有所进步，深入学习了NetworksX库的应用以及Gephi的操作方法。

本实验旨在通过可视化和可视分析技术，从选择的城市和房产中对不同贷款方式和

**关键词**：大数据可视化 复杂网络 文本处理

# 目 录

**摘 要**

[摘 要 3](#_Toc150809213)

[目 录 i](#_Toc150809214)

[第1章 引言 3](#_Toc150809215)

[1.1 实验背景 3](#_Toc150809216)

[1.2 实验内容 3](#_Toc150809217)

[1.3 实验环境 3](#_Toc150809218)

[第2章 模型设计 4](#_Toc150809219)

[2.1 文本描述 4](#_Toc150809220)

[2.2 文本处理 4](#_Toc150809221)

[2.3 网络建模 4](#_Toc150809222)

[2.4 网络可视化 4](#_Toc150809223)

[第3章 结果与讨论 5](#_Toc150809224)

[3.1 实验结果 5](#_Toc150809225)

[3.2 过程中遇到的困难 8](#_Toc150809226)

[参考文献 9](#_Toc150809227)

[个人评价 11](#_Toc150809228)

# 第1章 引言

## 1.1 实验背景

随着计算机硬件和大数据处理技术的高速发展, 海量数据智能分析的瓶颈已经从“如何快速地处理海量数据” 转变为 “如何从海量数据中快速有效地挖掘出有价值的信息”。可视化和可视分析基于人类的视觉感知特性, 结合数据分析和人机交互等技术, 利用可视化图表去解构复杂数据中蕴含的知识和规律。这种技术贯穿于数据科学的全生命周期, 被誉为大数据智能领域的最后一公里, 已在许多大数据应用分析场景取得令人瞩目的效果[1]。

## 1.2 实验内容

选择一本你喜欢的图书，如果是大部头多卷作品的，选其中一部即可，如《冰与火之歌》共发表5部，可选第1部《权力的游戏》；下载该图书txt文档；设计算法并用Python实现，提取出文字中的复杂网络模型，可以进行分词，虚词过滤，同义名词预处理等操作，以保证网络建模更加有效；调用NetworkX包，将建模的网络生成.gexf格式的网络数据，采用Gephi或NetworkX计算中心性、社团；导出中心性数据，使用Excel进行中心性分布可视化；导出社团数据，使用Excel进行网络中节点在不同社团比重的可视化；使用Gephi，对网络结构进行可视化；

通过可视化和统计数据，配图说明两个问题：1）哪一个你曾经对这本书中一个问题的感性认知，在可视化和统计分析过程中，找到了证据？2）哪一个你曾经对这本书的一个感性认知，在可视化和统计分析过程中，被推翻?

## 1.3 实验环境

处理器 AMD Ryzen 7 5800H with Radeon Graphics

内存 16 GB

系统 64 位操作系统, 基于 x64 的处理器

IDE PyCharm Professional 2023.2.3

# 第2章 模型设计

## 2.1 文本描述

《朝闻道》是中国科幻作家刘慈欣的中篇小说，收录在同名短篇集中。这部小说讲述了一群科学家为了探索宇宙的真理，不惜用自己的生命作为代价的故事。小说中充满了哲学意味和宗教意蕴，对于宇宙最深处真理的探寻，对于人类求知精神的赞颂，这些内容都是建立在科学的外壳之下，进行思维的碰撞和纠葛。

小说的题目《朝闻道》出自《论语·里仁》，子曰：“朝闻道，夕死可矣。”这句话表达了儒家对于“道”的崇高追求，把“道”的价值看得比生命还重。刘慈欣在这部小说中把儒家的“道”具象化为现代社会的科学真理，展现了科学家们为了真理而不惜牺牲的精神。小说的结尾，霍金出现，问排险者“宇宙的目的”是什么，把排险者问倒了，于是霍金没有变成火球被献祭而是继续活了下来。这一段显示了刘慈欣对于霍金的敬意，也暗示了宇宙的真理可能永远无法被人类完全揭示，只能不断地探索和质疑。

## 2.2 文本处理

在函数中使用jieba分词工具对文本进行分词，同时使用停用词列表过滤停用词。采用用户词典设置，如果在用户词典中输入的词都将被分作名字，避免出现名字被截断的现象。

在接收分词后的名字列表后，使用哈尔滨工业大学的LTP（Language Technology Platform）进行词性标注，LTP提供了一系列中文自然语言处理工具，用户可以使用这些工具对于中文文本进行分词、词性标注、句法分析等等工作。筛选出名字，形成名字列表。

## 2.3 网络建模

接收名字列表后，使用NetworkX库构建一个网络模型。节点表示名词，节点的权重表示名词的出现次数，边表示名词之间的共现关系，边的权重表示共现次数。每出现一对相邻的词，则无向边加一。

## 2.4 网络可视化

接收构建好的网络模型，利用Matplotlib库对网络进行可视化。节点的大小和颜色根据节点的权重进行映射，边的粗细和颜色根据边的权重进行映射。整个网络图使用Spring布局使得节点分布更加美观。接下来将构建好的网络模型写入一个.gexf格式的文件，以便后续在Gephi计算中心性数据和社团数据，并对数据进行可视化

# 第3章 结果与讨论

## 3.1 实验结果

对小说《朝闻道》完成文本处理和网络建模后导入Gephi，得到如下网络拓扑结构。

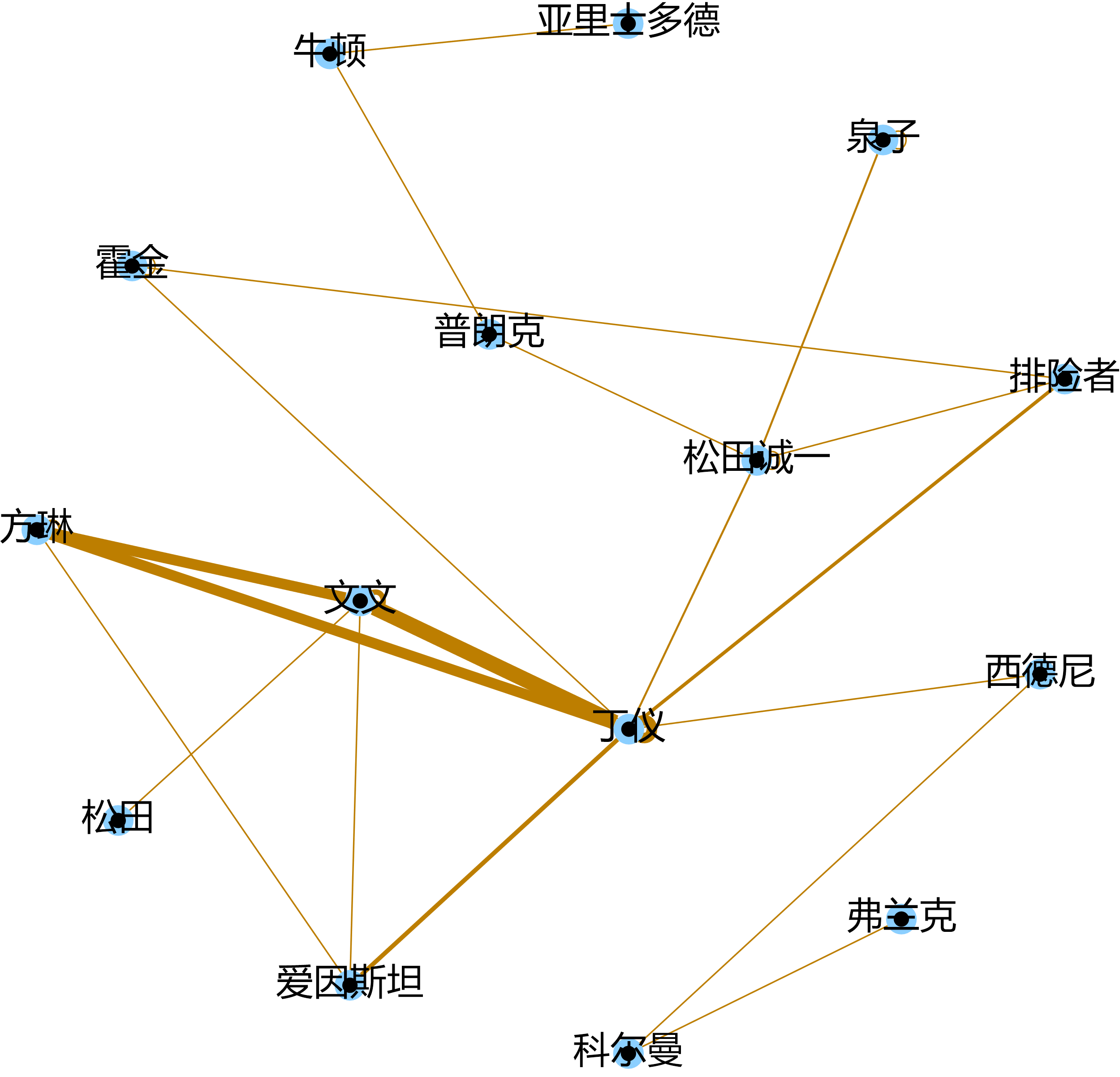


图1 《朝闻道》小说人物关系矢量图

从该网络结构中不难看出，小说主要是围绕丁仪、方琳和文文一家人进行情节展开，存在较强的中心化。同时丁仪与排险者的联系比其他人都强，也说明了丁仪在推动故事情节发展方面有着较大的作用。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **表1** | | | | |
| 网络图属性和计算结果 | | | | |
| Id | 标签 | 权重 | modularity\_class | closnesscentrality |
| 1 | 方琳 | 25 | 0 | 0.4 |
| 2 | 爱因斯坦 | 7 | 0 | 0.4 |
| 3 | 丁仪 | 66 | 0 | 0.56 |
| 10 | 松田 | 1 | 0 | 0.297872 |
| 12 | 文文 | 44 | 0 | 0.411765 |
| 0 | 排险者 | 30 | 1 | 0.4375 |
| 5 | 霍金 | 4 | 1 | 0.378378 |
| 7 | 牛顿 | 2 | 1 | 0.285714 |
| 8 | 普朗克 | 1 | 1 | 0.368421 |
| 9 | 泉子 | 4 | 1 | 0.333333 |
| 11 | 松田诚一 | 8 | 1 | 0.482759 |
| 14 | 亚里士多德 | 1 | 1 | 0.225806 |
| 4 | 弗兰克 | 1 | 2 | 0.241379 |
| 6 | 科尔曼 | 2 | 2 | 0.311111 |
| 13 | 西德尼 | 1 | 2 | 0.411765 |

表1为Gephi导出的点数据表，可以清晰看出人物被分为3个社团，并且分别具有不同的权重。

图2 《朝闻道》社团人数分布矢量图

图3 《朝闻道》人物权重

图4 《朝闻道》人物中心性

结合图3和图4来综合分析，可以发现丁仪、松田诚一、排险者具有较高的Closeness Centrality值，表明它们在网络中更接近其他节点，可能更能快速地与网络中的其他节点进行交互。同时可以关注到，在权重较高的人物中，大部分都具有相对较高的Closeness Centrality值，这与小说的内容是符合的。还可以看出，一般同一社团内的人物更偏向于具有相近的Closeness Centrality值，说明同一社团内的人物可能更偏向于同时出现。

在该网络中，每个社团内部的节点之间的连接相对紧密，但是各个社团的连接更为稀疏。

## 3.2 过程中遇到的困难

在该实验的过程中，我多次遇到人名识别的问题。最开始想要通过人工来进行人名标注，但是工作量比较大，之后发现LTP具有这个能力，所以开始学习NLP软件使用。但同样发现，该软件也无法较好地处理该问题，于是更换了模型来利用cuda的能力进行识别，最终得到比较好的效果。

在网络模型生成方面，由于之前不是很熟悉NetworkX的使用，在此遇到了较多Python语法方面的问题，后面进一步学习之后成功输出了图文件。

# 参考文献

1. Luo YY, Qin XD, Xie YP, Li GL. Intelligent Data Visualization Analysis Techniques: A Survey [J/OL]. Journal of Software (in Chinese), 1-49[2023-11-11].

(骆昱宇,秦雪迪,谢宇鹏等.智能数据可视分析技术综述[J/OL].软件学报:1-49[2023-11-11].https://doi.org/10.13328/j.cnki.jos.006911.)

1. Qin X, Luo Y, Tang N, et al. Making data visualization more efficient and effective: a survey[J]. The VLDB Journal, 2020, 29: 93-117. https://doi.org/10.1007/s00778-019-00588-3.

# 个人评价

本次实验的目的是解析一篇小说中的人物网络结构，利用Gephi进行可视化操作，并对社团和中心性进行数据分析。个人评价95分，在完成本次实验的过程中，有以下几个地方做的比较好：

在理论学习方面，我深入掌握了数据清洗、存储、解析的过程，针对人名识别学习了一定的NLP知识，同时对于网络模型的生成有了更深刻的理解，具备了一定的工程实践能力。后续可以根据类似要求进行其他项目的可视化工作。

在编程实现方面，我采用了模块化的编程思想，将每一个步骤封装成一个函数，然后在主程序中调用，实现了代码的高效和可维护。我还注意了代码的规范和注释，提高了代码的可读性和可理解性。

在软件学习方面，对于Gephi和Excel的可视化操作更熟悉了，具备实际投入使用的能力，相信这会对我以后的学习和工作产生非常大的帮助。

在文档撰写方面，按照要求包括了目录、中文摘要、关键词、小节分明的结构，以及背景信息、相关技术、模型或算法设计、实施和实验验证、实验分析、明确的结论和心得体会等。确保了实验过程全面呈现，读者可以轻松读懂实验思路。

总体来看，在本次实验中学习到了较多实用技能，提高了我的理论水平和实践能力，也增强了我的兴趣和信心。